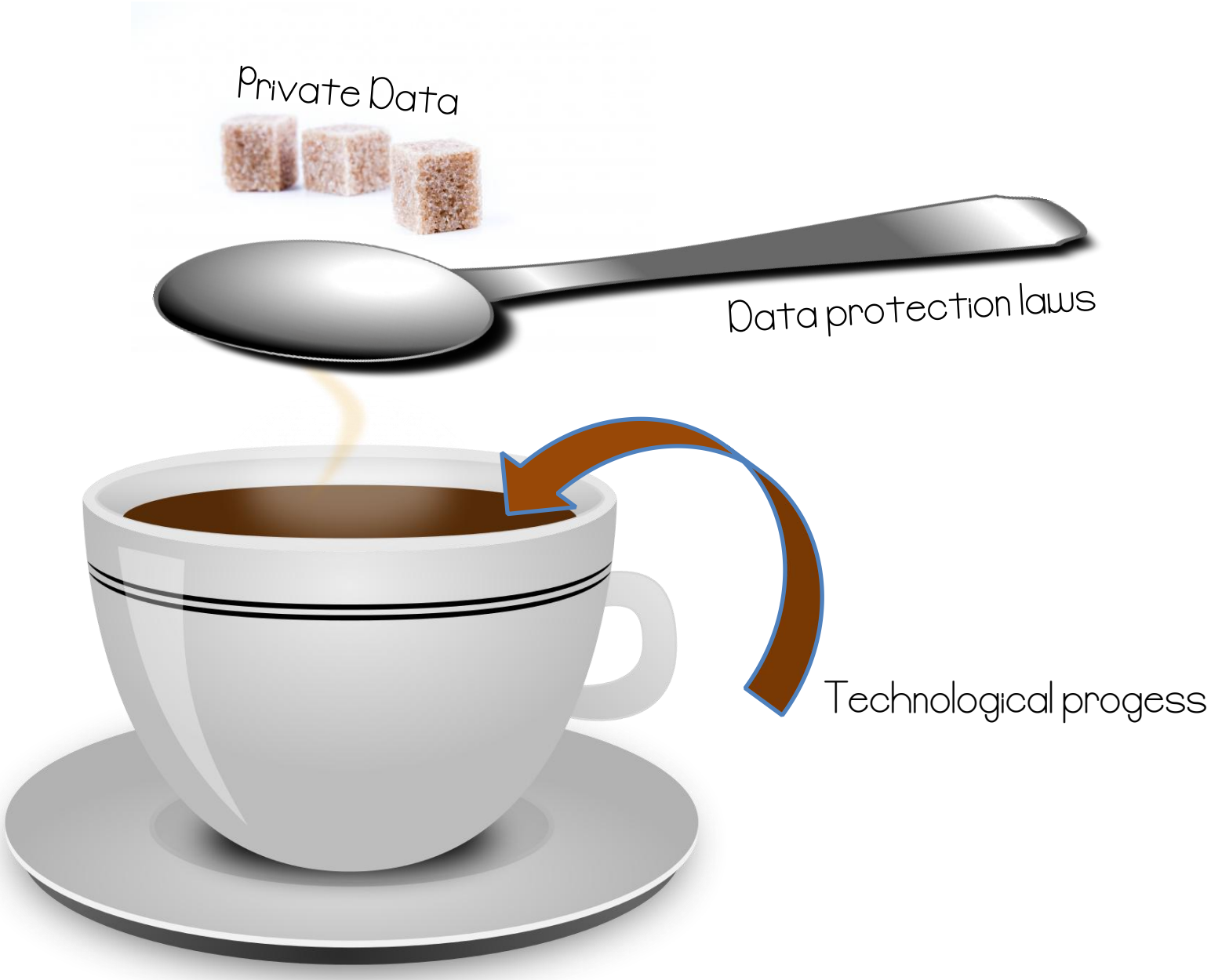
CD-MAKE 2017

DO NOT DISTURB?
Classifier behavior on perturbed datasets

Bernd Malle, Andreas Holzinger

b.malle@hci-kdd.org

Privacy in the 21st century… ??



GDPR - The right to be forgotten & more...

- Basically: A user has the right to have their data deleted from a database upon request
  - from which databases? backup? statistics? ML???
- The processing of personal data is expressely prohibited: age, race, income, socioeconomic status...
- BUT: algorithms must not be implicitly discriminating either!
- Think about recommender systems (collaborative filtering) ...
- From May 2018 onwards, the GDPR will come into force and violations carry a severe penalty (a few % of revenue...)

Different experimental scenarios

1. Simulate users exercising their “right to be forgotten” in the worst way possible – requesting the erasure of the **most valuable** data points in the knowledge base.

2. Anonymizing data in the first place and applying our classifiers on that anonymized datasets.

3. Perform a variant of 1) and delete "outliers" -> decrease variance in the dataset

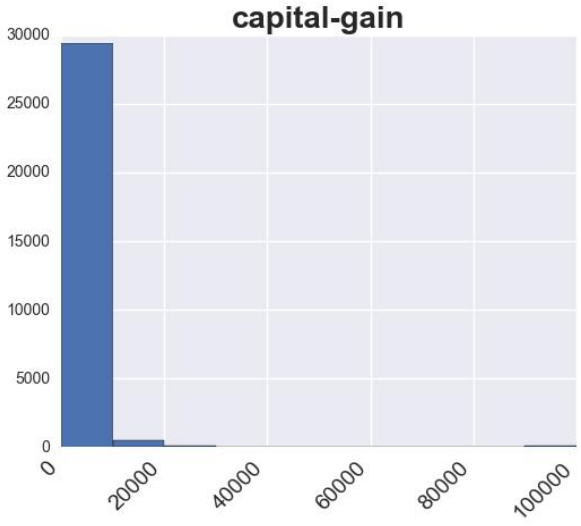
4. Combine outlier removal and anonymization (two detrimental forces at work)

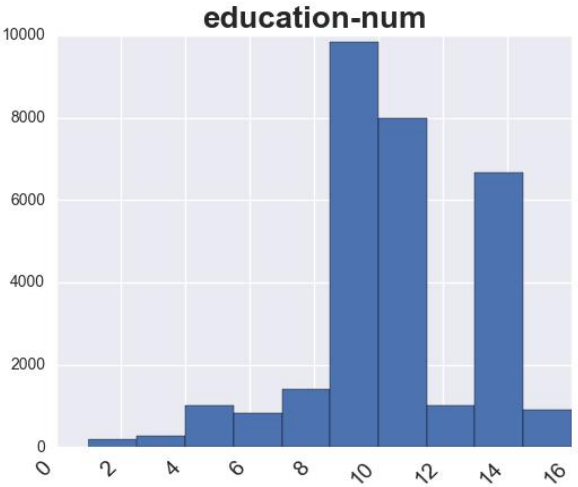
Scenario 1

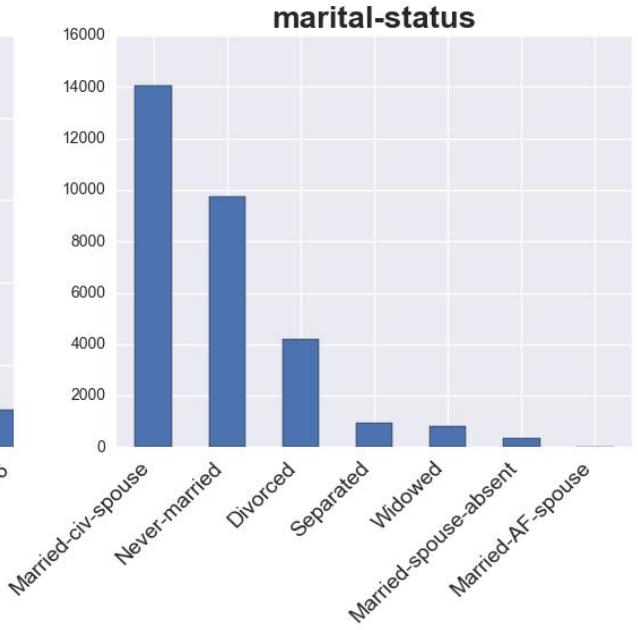
Selective deletion of valuable data points

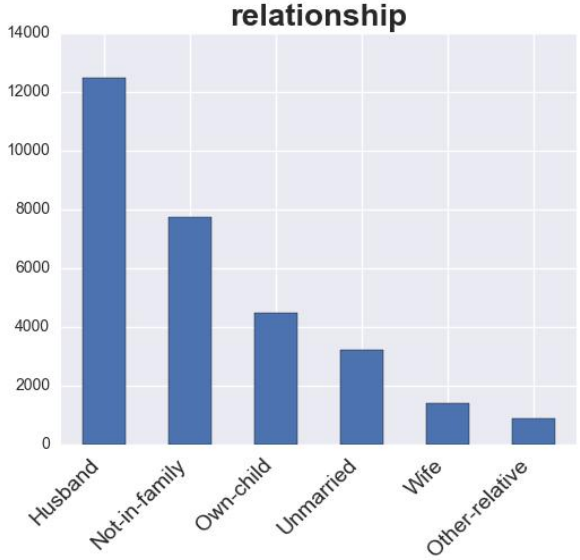
"simulating worst possible user behavior"

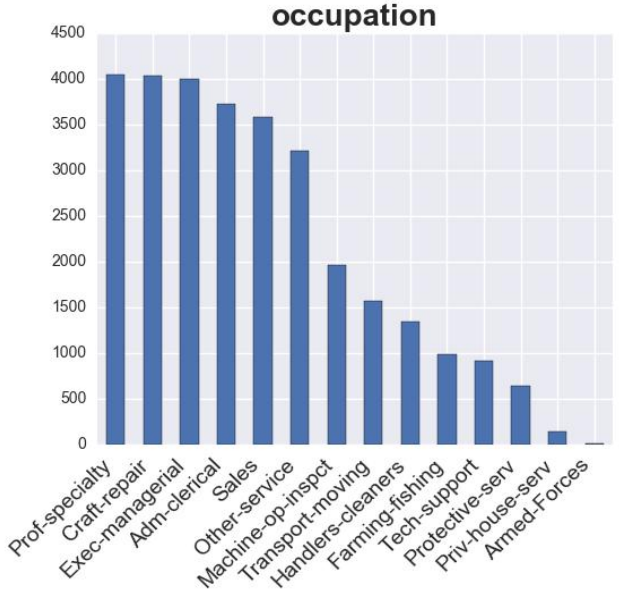
Adult dataset original distribution

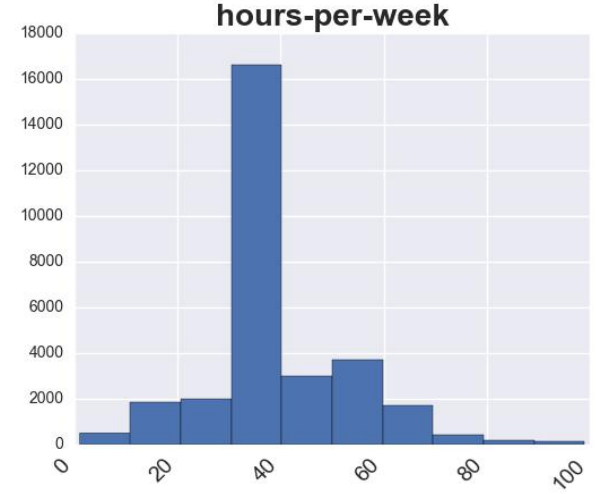






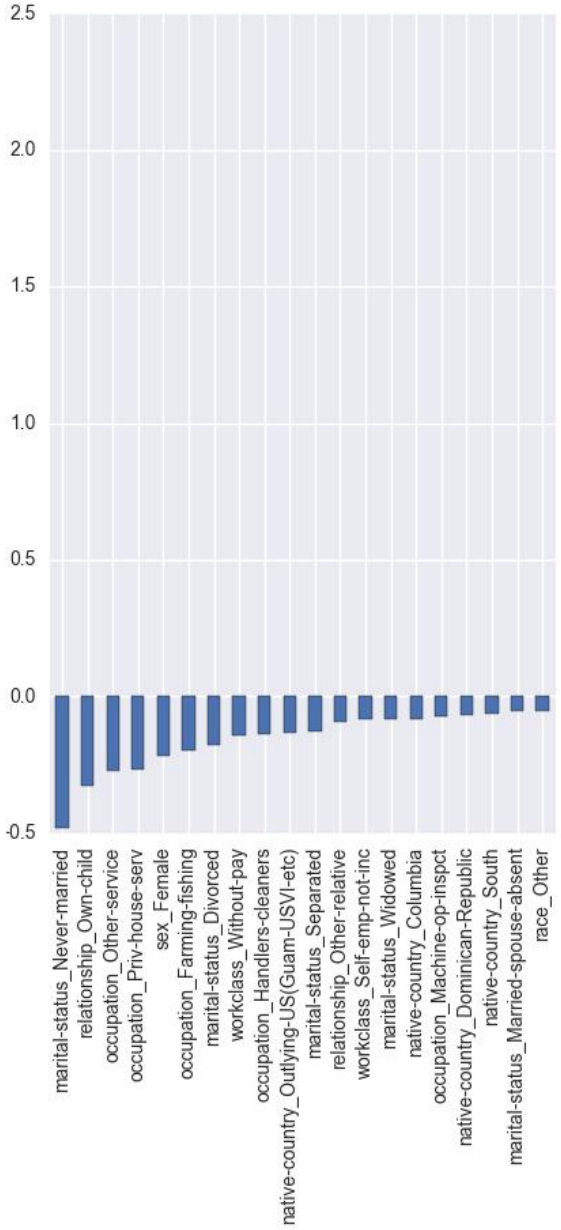


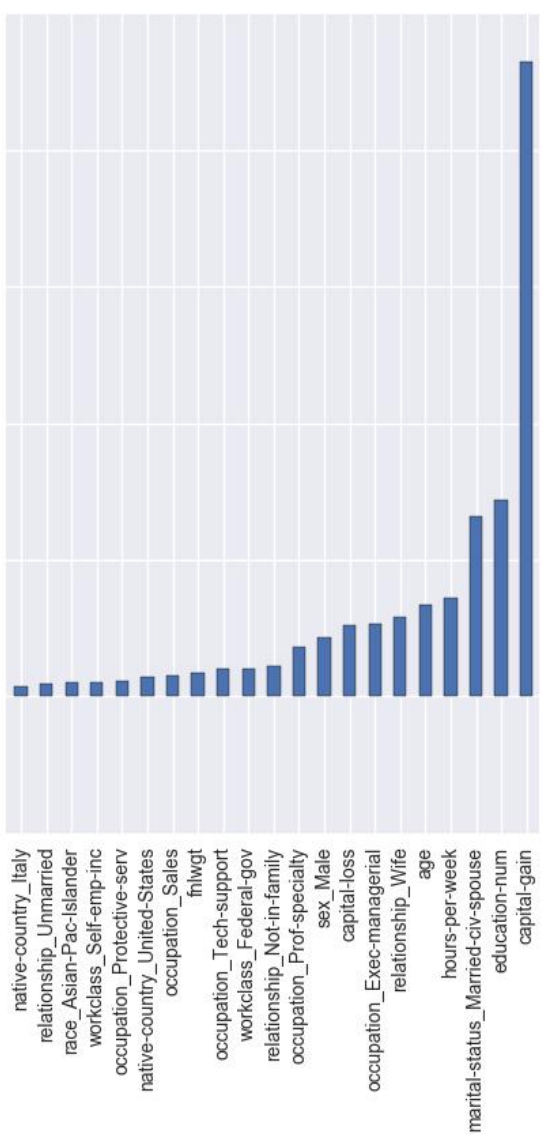




Find the most (in-)valuable data points

1. Train a logistic regression classifier
2. Retrieve the log coefficients
3. Highest coefs represent attribute values that contributed most to classifier certainty
4. Lowest were most confusing to the algorithm (noise?)

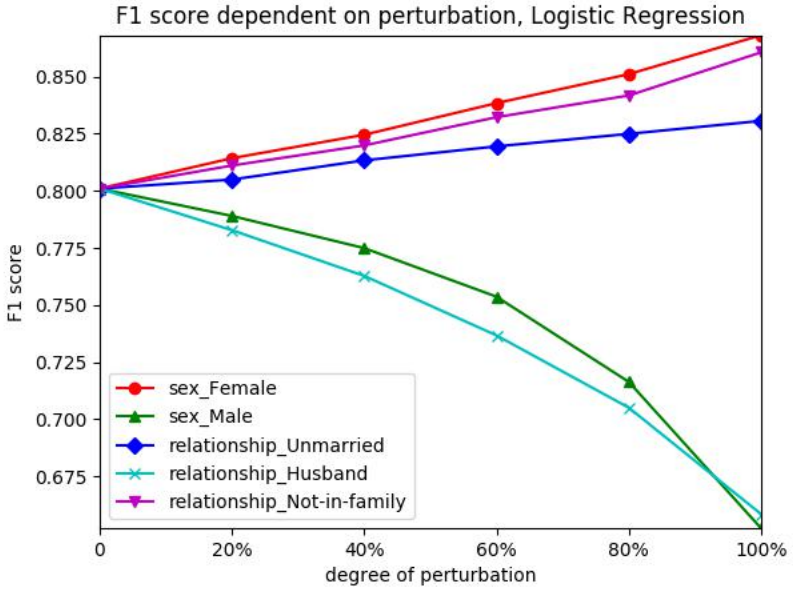


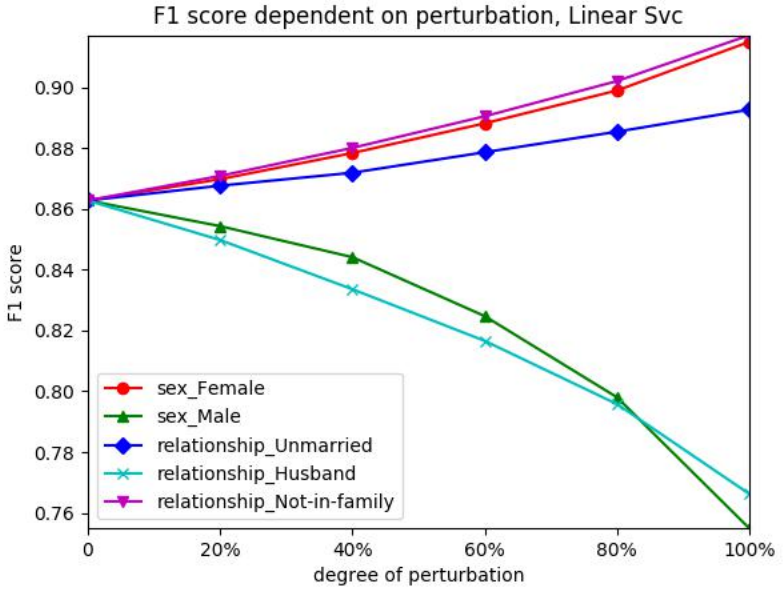


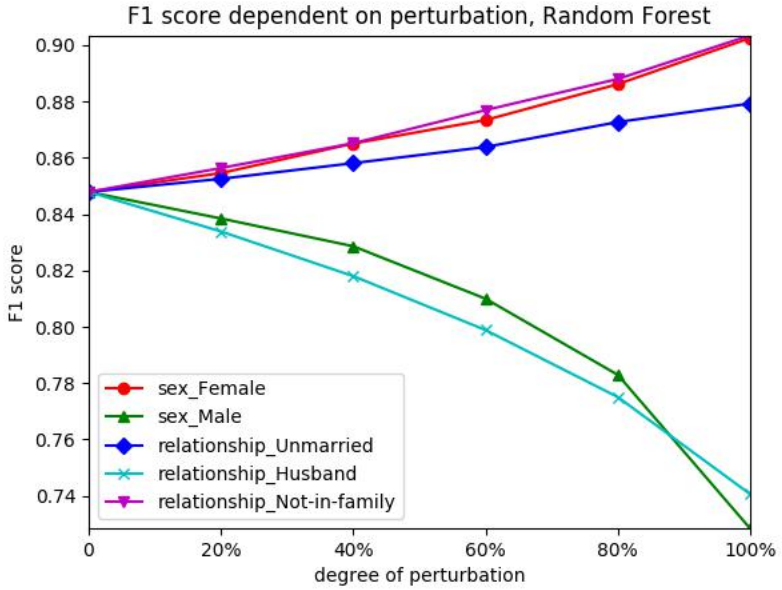
Selective deletion – prepared datasets

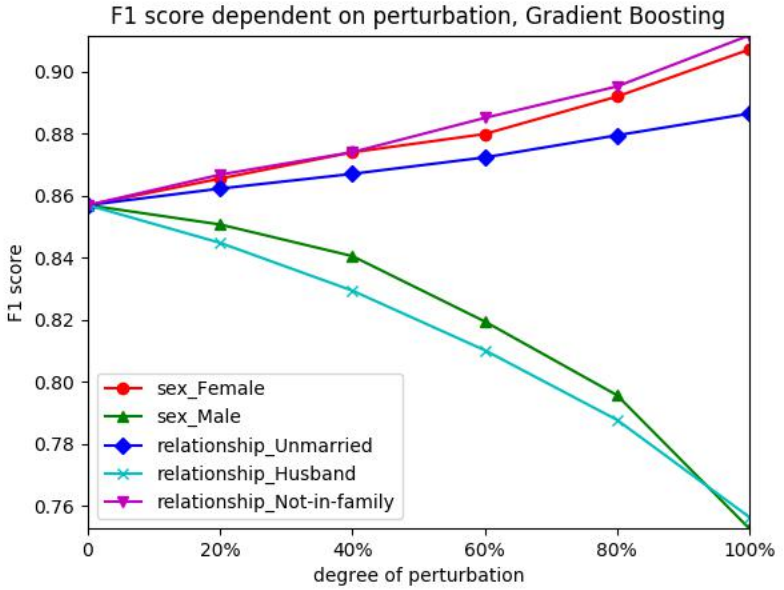
- We used multi-class classification with the targets
  - marital status => 7 categories
  - education_num => 14 continuous values, but clustered into 4 groups during preprocessing
- For each identified attribute value (occupation: *administrative clerk*) identified as valuable: delete 20%-100% of their containing rows in 5 steps
- This produced 5*#attributes new datasets per target
- We then used 4 classifiers on those data

Selective deletion - Results marital status

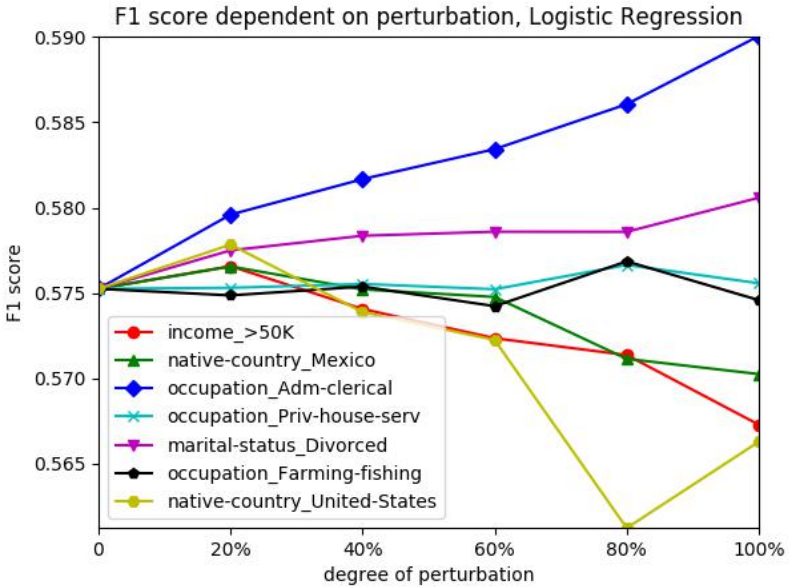


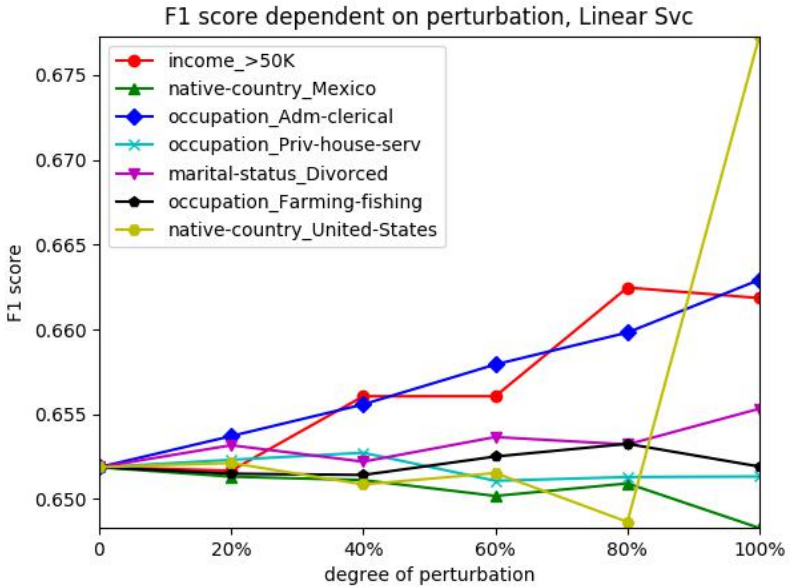


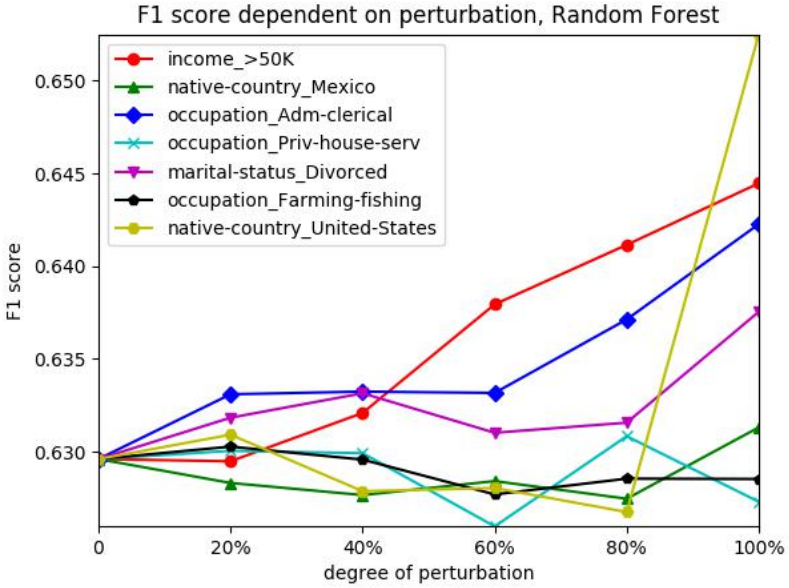


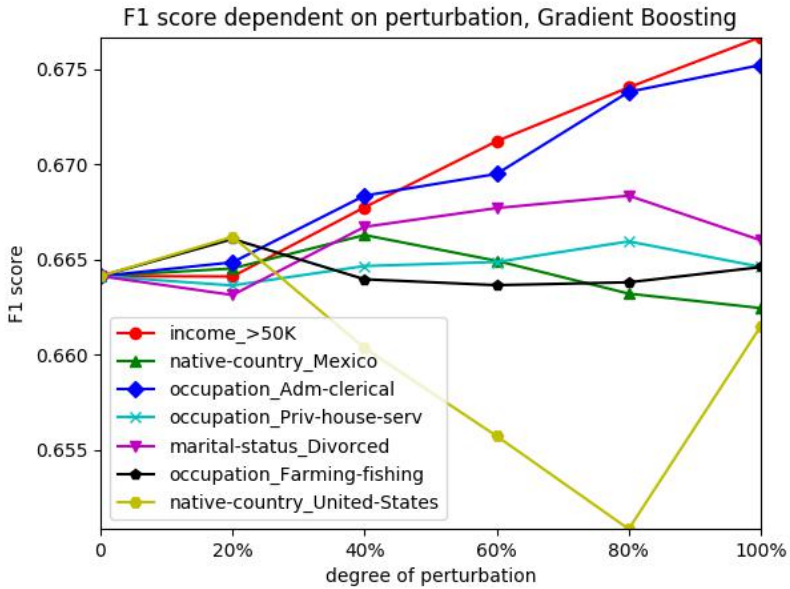


Selective deletion - Results education_num









Scenario 2

Wholesale anonymization of the knowledge base

"practical measurement of k-anonymization impact"

Bernd Malle <b.malle@hci-kdd.org>

Properties & General Approach

Data properties => Reduce granularity

Name	Age	Zip	Gender	Disease
Alex	25	41076	Male	Allergies
…	…	…	…	…

- Identifiers := immediately reveal identity
  - name, email, phone nr., SSN

  => DELETE

- Sensitive data
  - medical diagnosis, symptoms, drug intake, income

  => NECESSARY, KEEP

- Quasi-Identifiers := used in combination to retrieve identity
  - Age, zip, gender, race, profession, education

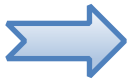
  => MAYBE USEFUL

  => MANIPULATE / GENERALIZE

K Anonymization criterion

k-anonymity: for every entry in the DS, there must be at least k-1 identical entries (w.r.t. QI's) => this is 3-anon:

Node	Name	Age	Zip	Gender	Disease
X1	Alex	25	41076	Male	Allergies
X2	Bob	25	41075	Male	Allergies
X3	Charlie	27	41076	Male	Allergies
X4	Dave	32	41099	Male	Diabetes
X5	Eva	27	41074	Female	Flu
X6	Dana	36	41099	Female	Gastritis
X7	George	30	41099	Male	Brain Tumor
X8	Lucas	28	41099	Male	Lung Cancer
X9	Laura	33	41075	Female	Alzheimer



Node	Age	Zip	Gender	Disease
X1	25-27	4107*	Male	Allergies
X2	25-27	4107*	Male	Allergies
X3	25-27	4107*	Male	Allergies
X4	30-36	41099	*	Diabetes
X5	27-33	410**	*	Flu
X6	30-36	41099	*	Gastritis
X7	30-36	41099	*	Brain Tumor
X8	27-33	410**	*	Lung Cancer
X9	27-33	410**	*	Alzheimer

Generalization hierarchies

- Generalization (hierarchies)
  - fixed ruleset
  - range partitioning (numerical values...)

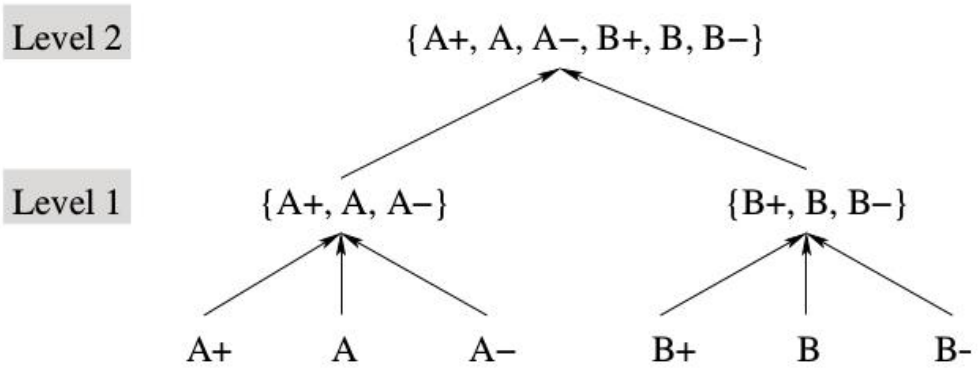
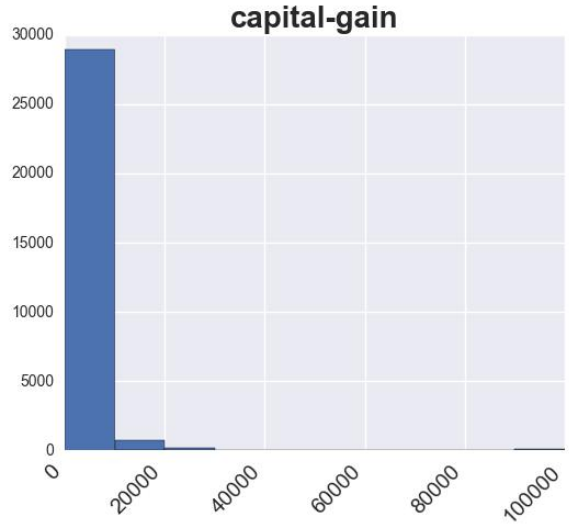


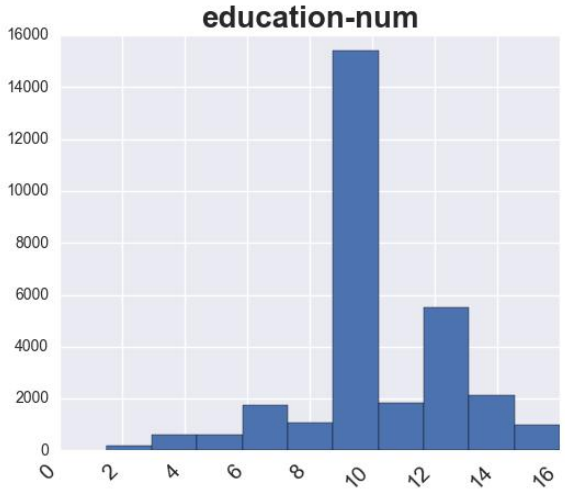
Figure 1: A possible generalization hierarchy for the attribute "Quality".

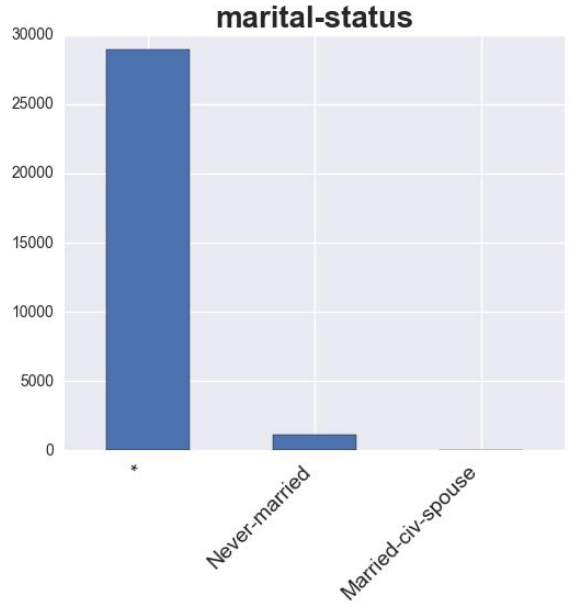
- Suppression
  - Special case of generalization (with one level)

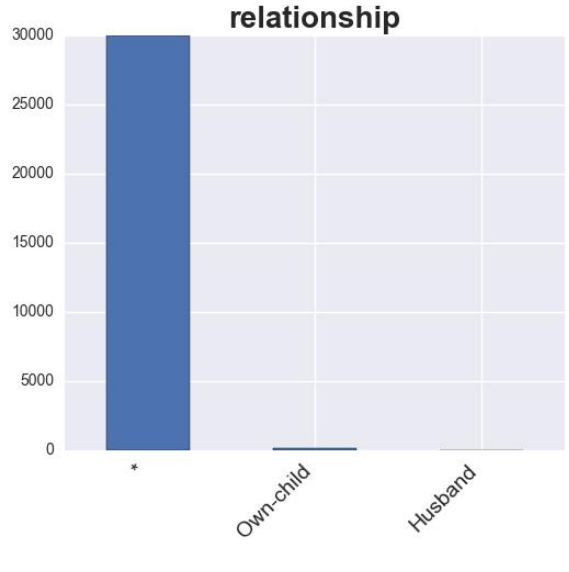
Graphics Source: Bayardo, R. J., & Agrawal, R. (2005, April). Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on* (pp. 217-228). IEEE.

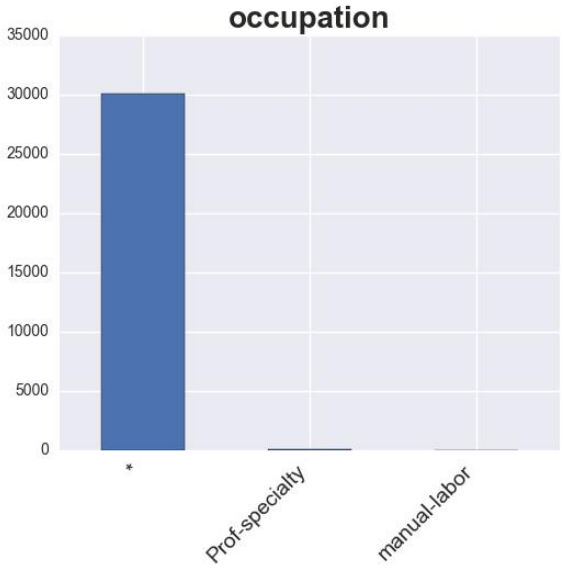
Anonymization - datasets

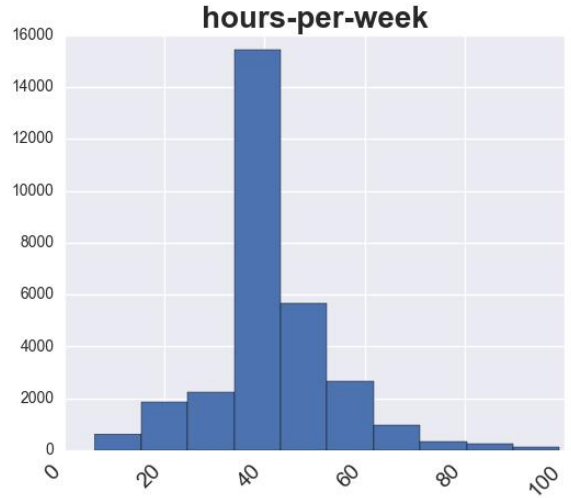












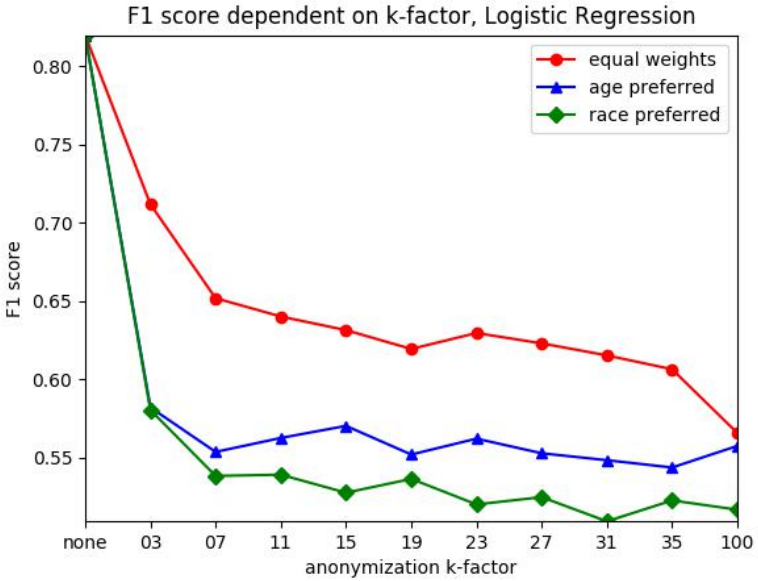
Anonymization – prepared datasets

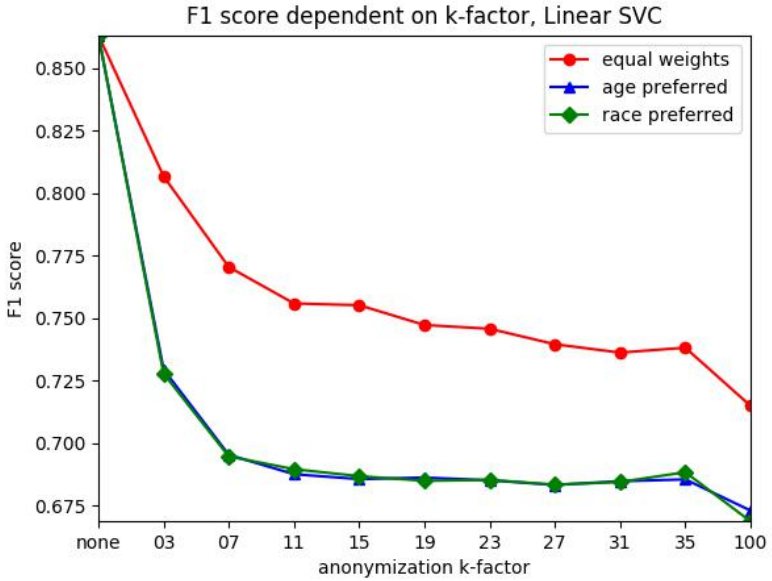
- We used k-factors of:
- range(3; 35; 4) as well as 100
- Each combined with three different weight vectors

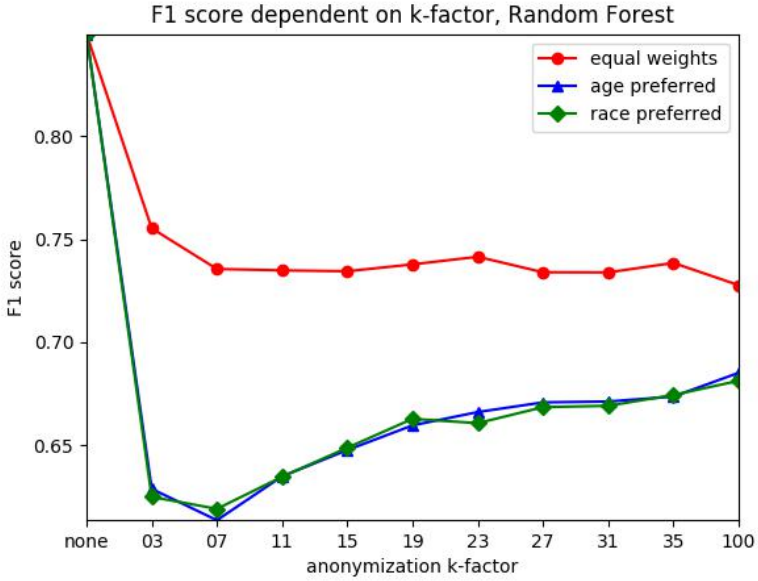
  Equal weights for all columns
  Age preferred (0.88 vs 0.01 rest)
  Race preferred (0.88 vs. 0.01 rest)

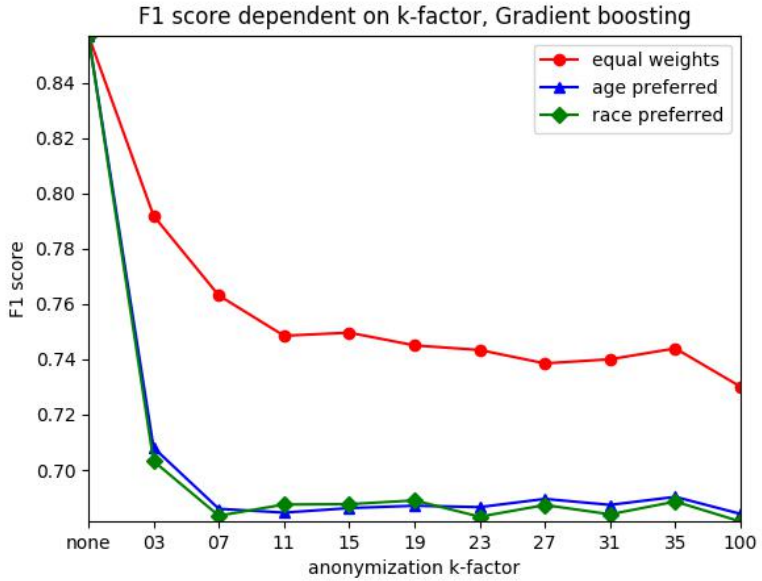
- Resulting in 15 differently anonymized data sets

ML on Anonymization - Marital Status

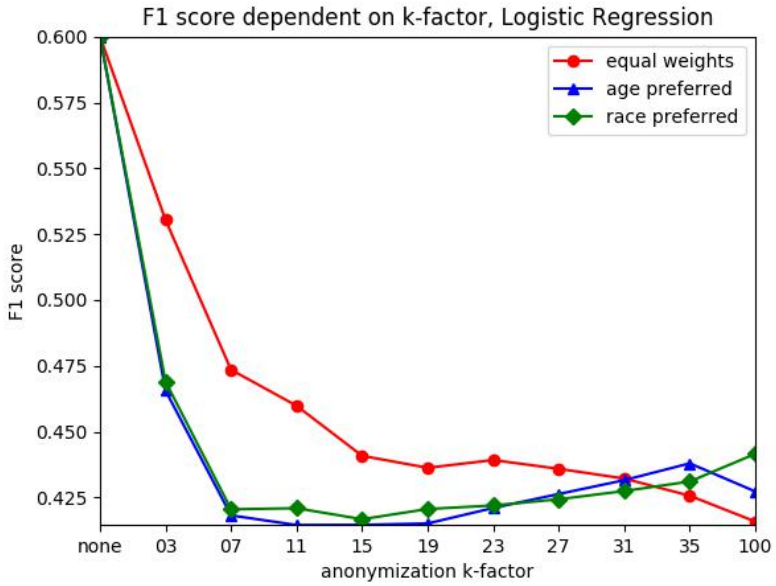


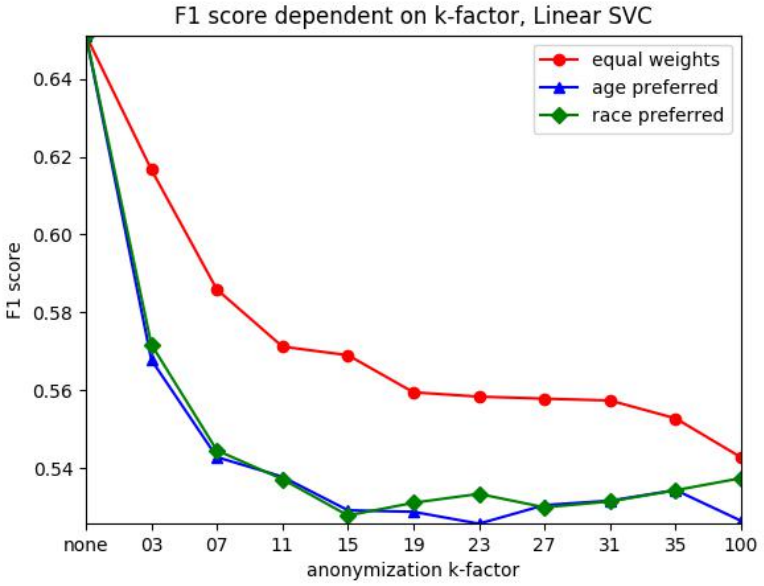


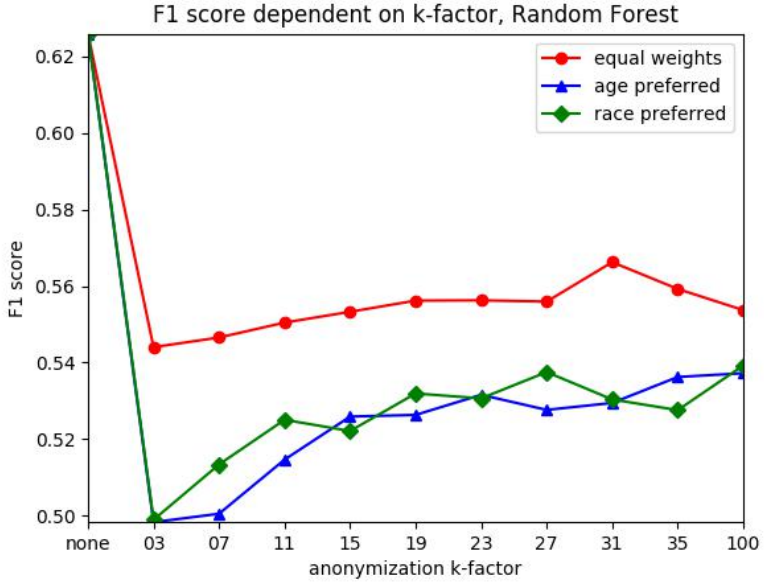


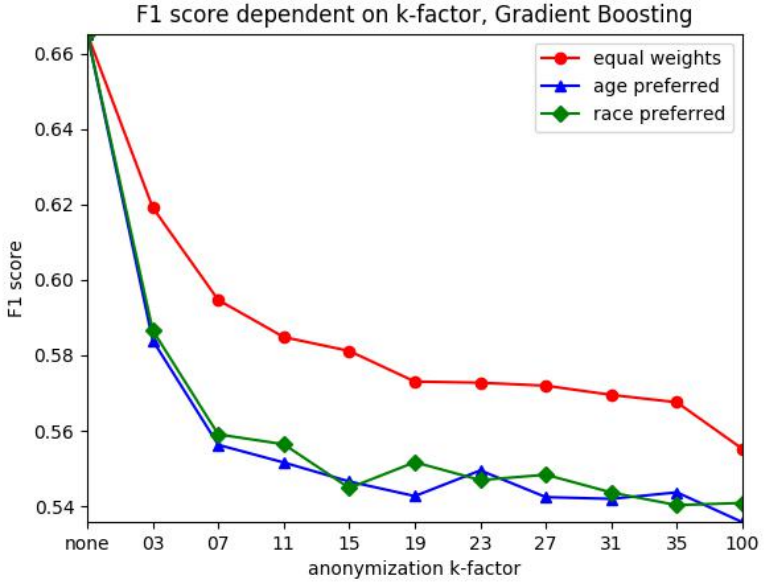


ML on Anonymization - Education Num









Scenario 3

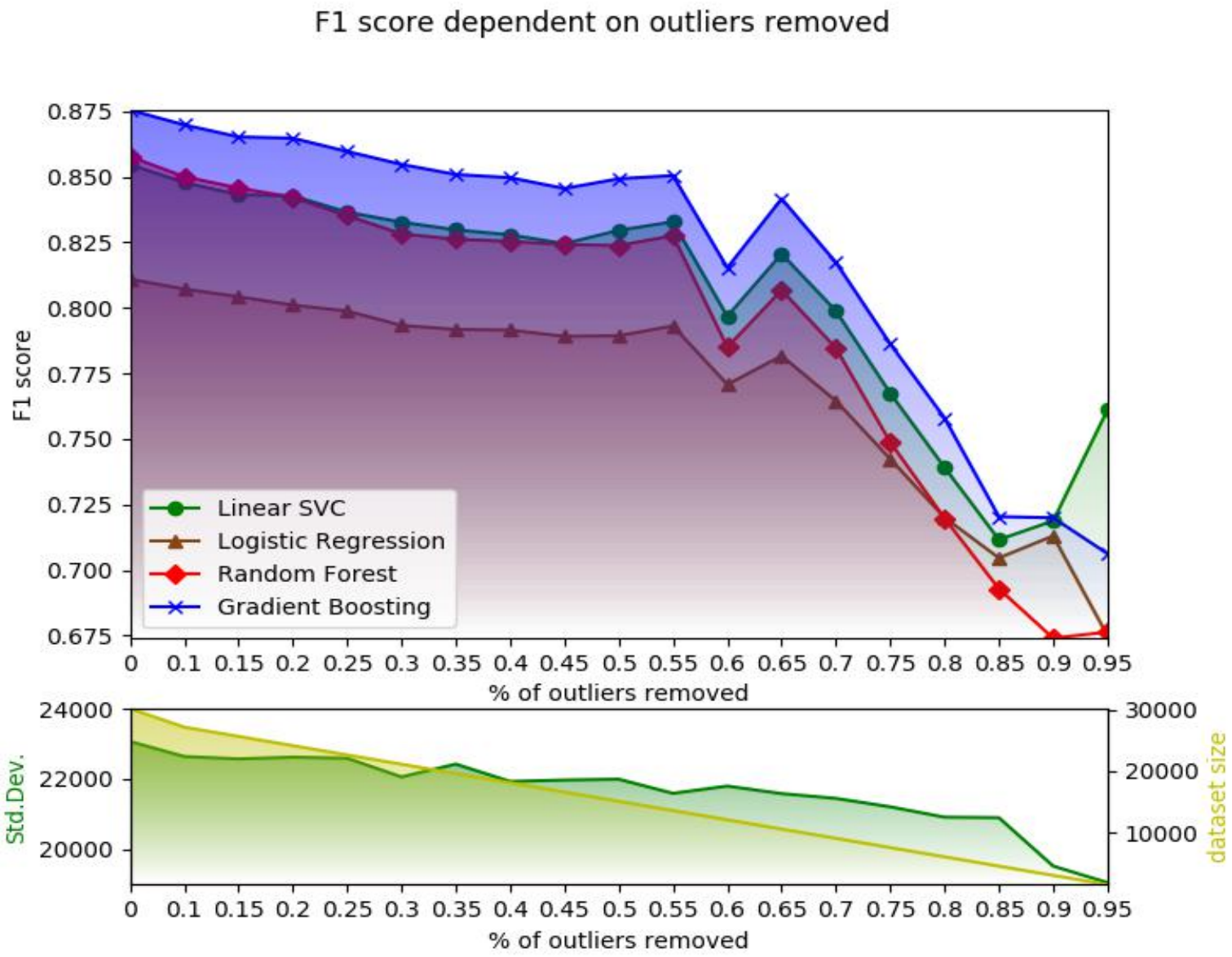
Outlier removal

"criminals / hate speech removal"

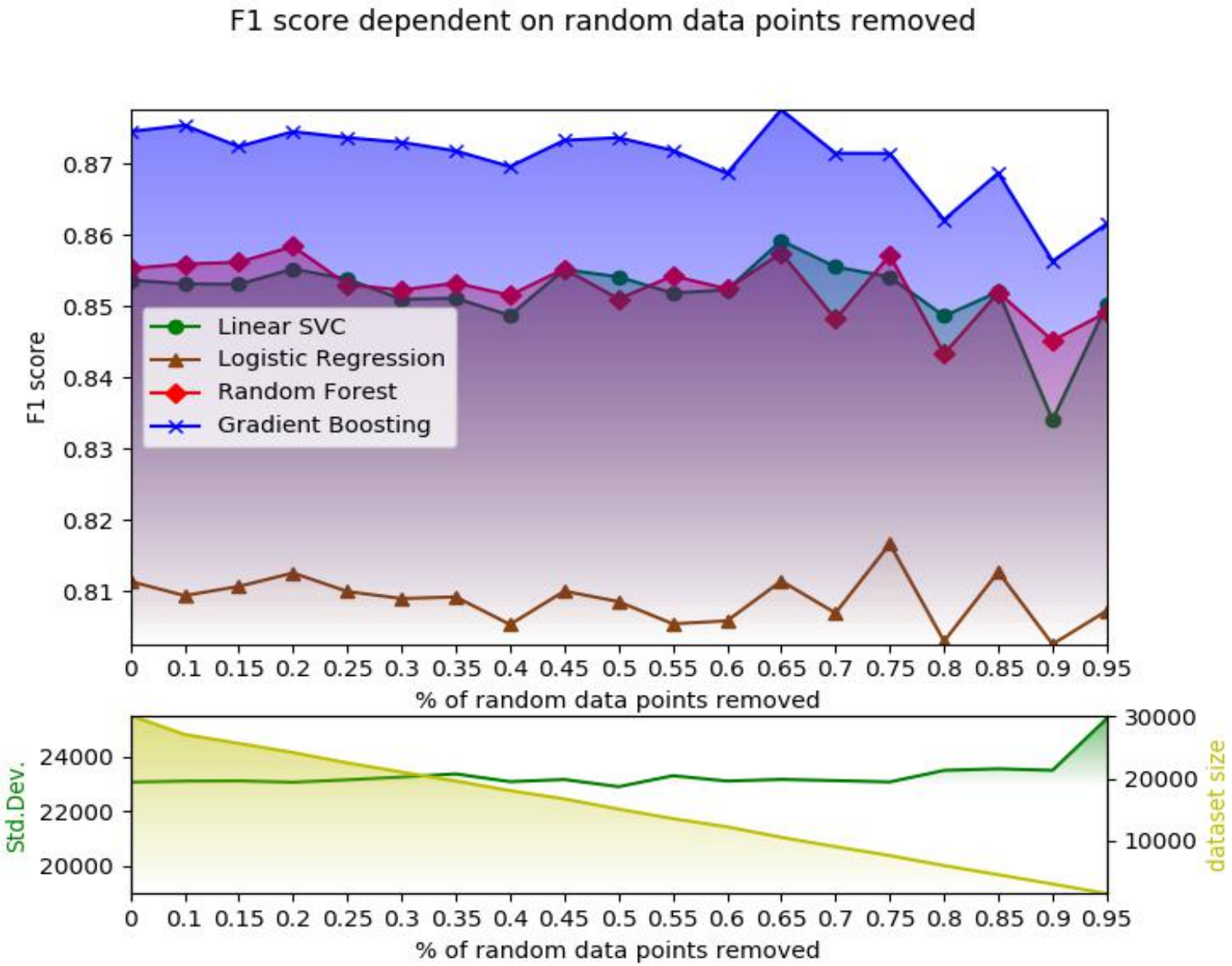
Outlier removal – prepared datasets

- This time we used income as target attribute (just for comparison to our earlier work..)
- "outlier" range(5; 95; 5)
- Of course such a range of outliers is simply a mutilation of the data set in order to decrease variance
- How do our classifiers perform with data becoming a more and more homogeneous lump of points?
- We also compared it to random deletion over the same range (to make sure it's not just the data set size...)

Outlier removal



Random deletion



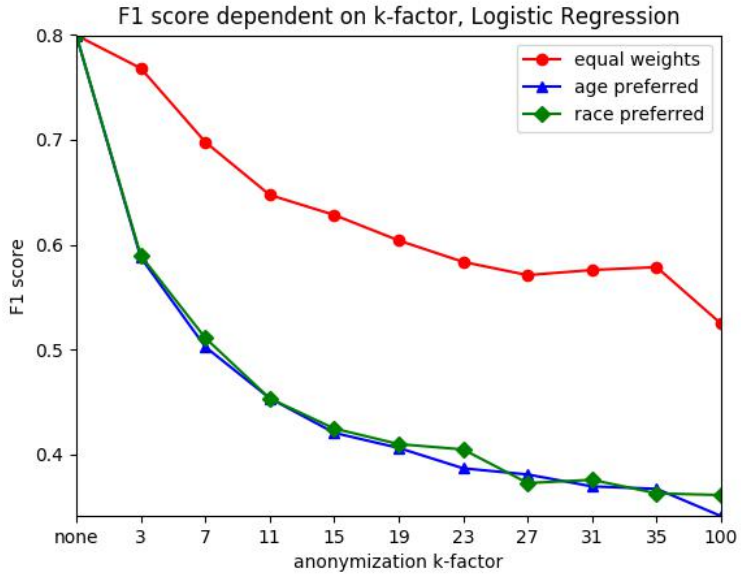
Scenario 4

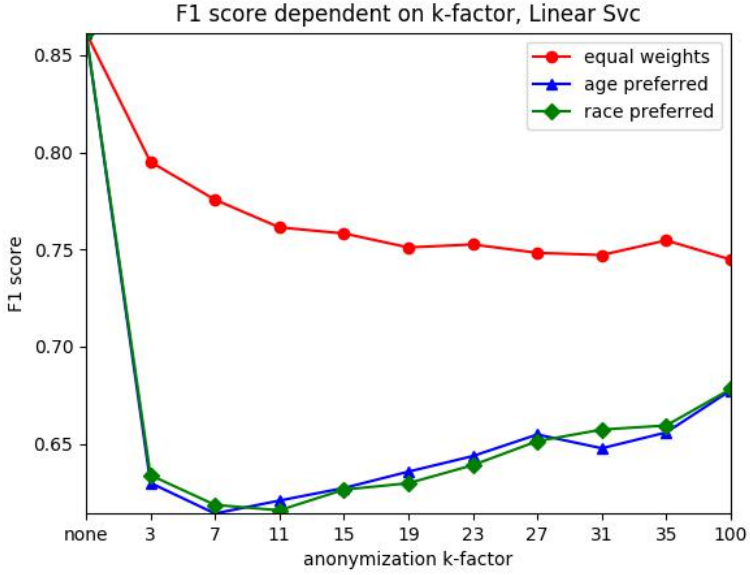
Outlier removal -> then anonymization

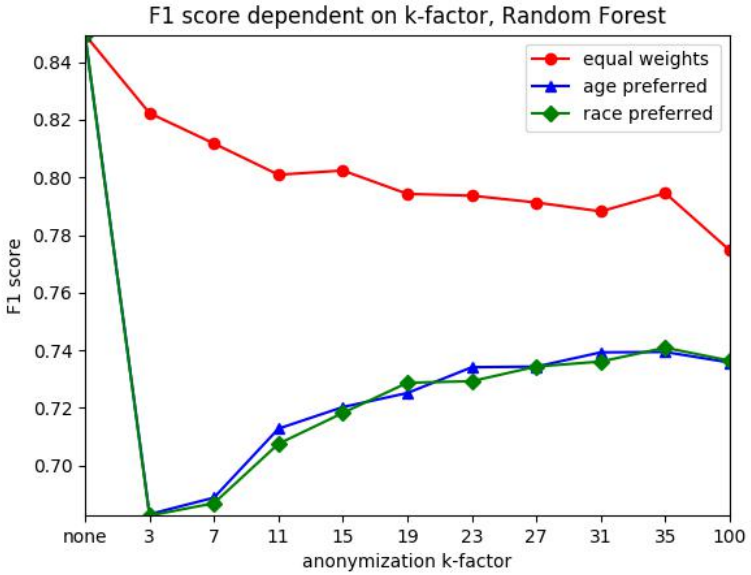
Outlier removal + anonymization - datasets

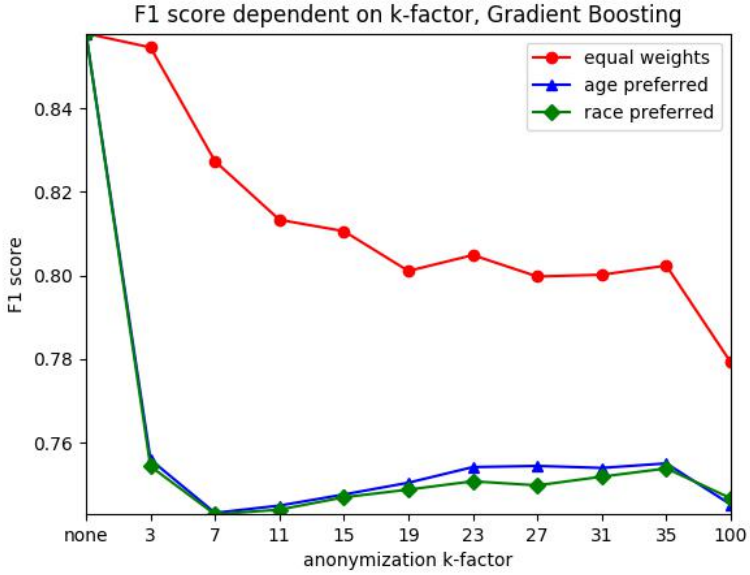
- Outliers decrease variance => make it harder for an algorithm to discriminate
- But more homogeneous data allow algorithms to find shallower generalization levels => less information loss during anonymization
- What will those two detrimental forces produce?
- scenario this time: "outlier" removal fixed at 30%
- then anonymization like before (k=3..35, 100)

ML on Anonymization on outliers removed









Lessons Learned

1. Succumbing to the “right-to-be-forgotten” still seems better than performing ML on (crudely) anonymized DBs

2. But => Less variance in the data-set makes it harder for algorithms to discriminate (and computationally costlier)

3. ML on anonymization on outlier removal produced surprisingly good results in the case of Gradient Boost & Logistic Regression

4. We (...somebody) should verify this on a multitude of interesting datasets, anonymization algorithms etc.

Future Work

1. Different forms of ML => Prediction, Clustering, Pattern Recognition, etc.

2. Different forms of anonymization => l,t,delta, perturbation (eps-diff-priv), micro-aggregation

3. Experiments on different data structures => especially graphs (social networks) would be interesting and highly relevant in our age

4. We (...somebody) should verify this on a multitude of interesting datasets, anonymization algorithms etc.

Thank you!